Appendix XVIII to Part 86—Statistical Outlier Identification Procedure for Light-Duty Vehicles and Light Light-Duty Trucks Certifying to the Provisions of Part 86, Subpart R

Residual normal deviates to indicate outliers are used routinely and usefully in analyzing regression data, but suffer theoretical deficiencies if statistical significance tests are required. Consequently, the procedure for testing for outliers outlined by Snedecor and Cochran, 6th ed., *Statistical Methods*, PP. 157–158, will be used. The method will be described generally, then by appropriate formulae, and finally a numerical example will be given.

(a) Linearity is assumed (as in the rest of the deterioration factor calculation procedure), and each contaminant is treated separately. The procedure is as follows:

(1) Calculate the deterioration factor regression as usual, and determine the largest residual in absolute value. Then recalculate the regression with the suspected outlier omitted. From the new regression line calculate the residual at the deleted point, denoted as $(y_i - y_i')$. Obtain a statistic by dividing $(y_i - y_i')$ by the square root of the estimated variance of $(y_i - y_i')$. Find the tailed probability, p, from the t-distribution corresponding to the quotient (double-tailed), with n-3 degrees of freedom, with n the original sample size.

(2) This probability, p, assumes the suspected outlier is randomly selected, which is not true. Therefore, the outlier will be rejected only if $1 - (1\text{-}p)^n < 0.05$.

(3) The procedure will be repeated for each contaminant individually until the above procedure indicates no outliers are present.

(4) When an outlier is found, the vehicle test-log will be examined. If an unusual vehicle malfunction is indicated, data for all contaminants at that test-point will be rejected; otherwise, only the identified outlier will be omitted in calculating the deterioration factor.

(b) Procedure for the calculation of the t-Statistic for Deterioration Data Outlier Test.

(1) Given a set of n points, (x_1, y_1), (x_2, y_2) * * * (x_n, y_n).

Where:

x_i is the mileage of the i^{th} data point.
y_i is the emission of the i^{th} data point.
Assume model:

$$y = a + \beta(x - \bar{x}) + \in$$

(2)(i) Calculate the regression line.

$$\hat{y} = a + b(x - \bar{x})$$

(ii) Suppose the absolute value of the i^{th} residual

$(y_i - \hat{y}_i)$ is the largest.

(3)(i) Calculate the regression line with the i^{th} point deleted.

$$\hat{y}' = a' + b'(x - \bar{x})$$

(ii)

er06jn97.004

Where:

y_1 is the observed suspected outlier.
$\hat{y}'_i$ is the predicted value with the suspected outlier deleted.

G:\GRAPHICS\ER06JN97.005

(x is calculated without the suspected outlier)

G:\GRAPHICS\ER06JN97.006

(iii) Find p from the t-statistic table

Where:

$p = \text{prob}\ (\mid t(n\text{-}3) \mid \geq t)$
$t(n–3)$ is a t-distributed variable with n–3 degrees of freedom.

(iv) y_i is an outlier if $1–(1–p)^n < .05$

x	y	ŷ	y − ŷ
8	59	56.14	2.86
6	58	58.17	−0.17
11	56	53.10	2.90
22[1]	53	41.96	11.04
14	50	50.06	−0.06
17	45	47.03	−2.03
18	43	46.01	−3.01
24	42	39.94	2.06
19	39	45.00	−6.00
23	38	40.95	−2.95
26	30	37.91	−7.91
40	27	23.73	3.27

[1] Suspected outlier.

(4)(i) Assume model:

$$y = a + \beta(x - \bar{x}) + \in$$
$$y = 45 - 1.013(x - \bar{x})$$

(ii) Suspected point out of regression:

$$y = 44.273 - 1.053(x - \bar{x})$$
$$y = 44.273 - 1.053(22 - 18.727) = 40.827$$
$$y_i - \hat{y}'_i = 12.173$$

G:\GRAPHICS\ER06JN97.007

[62 FR 31270, June 6, 1997; 62 FR 45289, Aug. 26, 1997, as amended at 63 FR 987, Jan. 7, 1998]